Modeling Retail Store Performances with Physical and Demographic Attributes

Statistical Learning Final Project

ERDEM BAKIR

65288A

Data Science for Economics, University of Milan

ABSTRACT

This research analyzes the primary factors influencing monthly sales revenue in the retail industry by utilizing a comprehensive dataset of 1,650 observations. The dataset is available on Kaggle with name of Retail Store Performance repository. The study follows a multi-step analytical framework: first, K-Means Clustering was applied to categorize stores into three distinct groups based on their operational profiles. Following this, Linear Regression and Random Forest algorithms were developed to compare their predictive performance.The results indicate that the Random Forest model achieved a high level of accuracy, with an R-squared of 0.805 and an RMSE of 29.422 on the validation set. This demonstrates that the model is reliable and can make consistent predictions on new data. According to the Variable Importance analysis, Product Variety (100.00) and Store Size (99.9) are the most significant drivers of sales, whereas factors like marketing budget and location have a much smaller impact. These findings suggest that retail managers should focus on expanding product diversity and store capacity to increase revenue. The study provides a solid evidence-based framework for making strategic decisions regarding operational planning and new store investments.

Contents

1 Introduction 1

2 Dataset 2

3 Preprocessing 4

3.1 Missing Values 4

3.2 Outlier Analysis 6

3.3 Data Transforming 6

3.4 Scaling 8

3.5 Correlations and Multicolinearity 9

4 Unsupervised Learning 11

4.1 K Means Clustering 11

4.1.1 Elbow Method and Silhouette for Optimal K 11

4.2 Creating Clusters 13

5 Supervised Learning 15

5.1 Linear Regression 15

5.2 Random Forest 18

6 Conclusion 20

Chapter 1

Introduction

Even though the usage of e-commerce marketplaces is increasing today, physical stores still play a crucial role in customer experience and revenue generation in the retail sector. In this dynamic and changing environment, data-driven location selection and performance optimization are essential for retail stores, rather than opening new locations randomly.

Distributors and companies need to analyze the performance of stores to allocate their resources most efficiently. However, every store varies in terms of location, physical space, competitiveness, and operational proficiency.

The primary aim of this work is to identify the crucial factors influencing store performance by using data from 1650 different store points. In this study, potential store revenues will be estimated using Supervised Learning methods. Additionally, through Unsupervised Learning methods, stores with similar attributes will be grouped and evaluated from different perspectives based on their performance.

Consequently, the results of this work will not only guide existing stores in making improvements but also serve as a strategic tool for investors deciding where to open new stores and for companies in their product distribution planning.

Chapter 2

Dataset

The data set of this work contains monthly performance of 1650 different retail store points. The data set has 13 features like physical features, operational efficiencies, product variation, and marketing spending of stores.

The focus of this study is to predict the target variable, Monthly Sales Revenue, using independent variables that potentially shape the financial output. We hypothesize that a larger Store Size and greater Product Variety will lead to an increase in revenue, as customers will have a more comprehensive shopping experience with more space and product options.

A similar positive relationship is also expected with operational variables such as Marketing Spend and Promotions Count. An increase in these variables is likely to boost Customer Footfall the number of people visiting the store and consequently, total revenue will rise. Additionally, Employee Efficiency represents the service quality of the store, while Store Age reflects the effects of customer loyalty.

While considering store related features, we must also consider competitors seeking market share in the sector. Our expectation is that a higher Competitor Distance results in higher revenue, as competition decreases. However, there is a small chance that competition sometimes increases revenue in both stores due to market agglomeration. Finally, socio-economic and geographical features like Store Location and Economic Indicator will reveal the relationship between consumer spending and the economic situation, highlighting the importance of location. Specifically, the categorical feature StoreLocation covering four major cities: Los Angeles, Sacramento, Palo Alto, and San Francisco. Similarly, the StoreCategory variable classifies the retail outlets into three distinct sectors: Electronics, Grocery, and Clothing, allowing for an analysis of performance across different market segments.

Table 2.1: Dataset Variable Descriptions and Characteristics

Variable	**Type**	**Description**
MonthlySalesRevenue	Num.	Total monthly sales revenue generated by the store.
StoreSize	Num.	Physical floor area of the store (m^2).
StoreLocation	Cat.	Name of city the store is located.
StoreCategory	Cat.	Type of store format (e.g., Supermarket).
CompetitorDistance	Num.	Distance to the nearest competitor store (km).
MarketingSpend	Num.	Monthly marketing expenditure for the specific store.
CustomerFootfall	Num.	Average monthly number of customer visits.
EmployeeEfficiency	Num.	Efficiency score of the staff (0-100).
ProductVariety	Num.	Total number of different product SKUs available.
PromotionsCount	Num.	Number of active promotional campaigns in the month.
StoreAge	Num.	Number of years the store has been in operation.
EconomicIndicator	Num.	Regional economic prosperity index.

Note: Categorical variables will be transformed using One-Hot Encoding.

Chapter 3

Preprocessing

3.1 Missing Values

Missing value analysis was performed using standard checks, and no imputation was necessary as the data contained no NA or null entries.

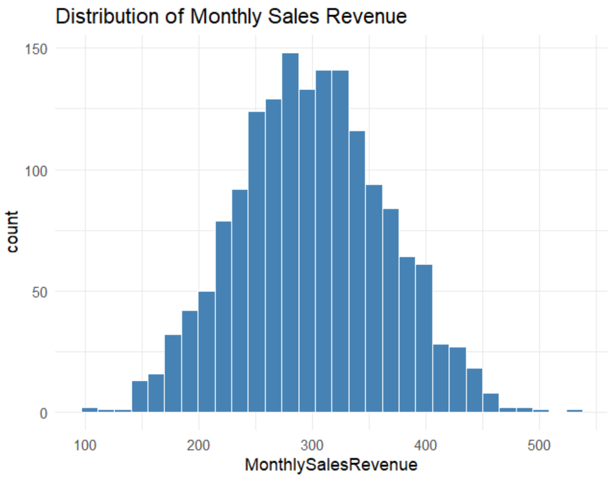

Figure 3.1: Distribution of Target Variable, Monthly Revenue Sales

After checking distribution of target variable, it is observed that it is distributed normally, which doesn't need any process about it.

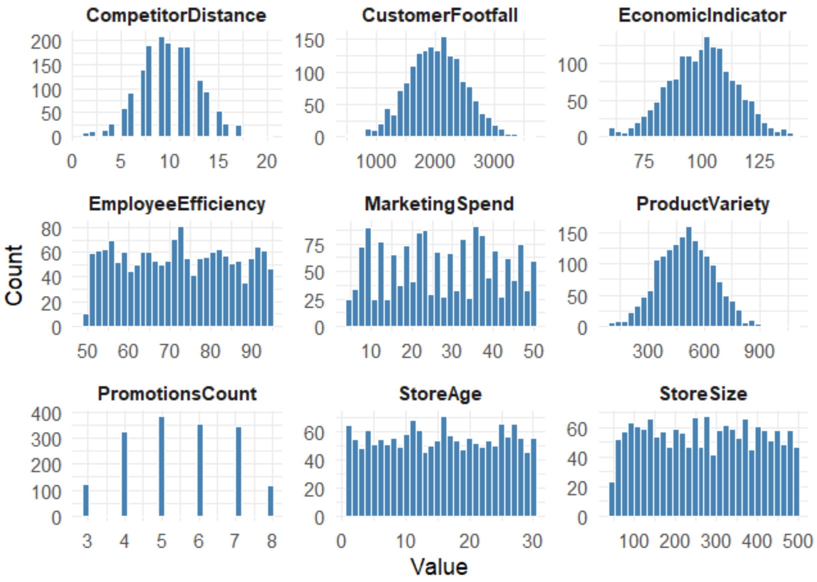

Figure 3.2: Distribution of All Numerical Independent Variables

As we can see from the histograms, the variables follow two main patterns.

First, variables like CustomerFootfall, ProductVariety, and EconomicIndicator show a Normal Distribution (bell shape). This means that most of the data is close to the average, which is very useful for our linear regression model.

Second, variables like StoreSize and StoreAge look flat, which is called a Uniform Distribution. This shows that our dataset is balanced. For example, we have an equal number of small and large stores, so the data is not biased towards one specific group.

3.2 Outlier Analysis

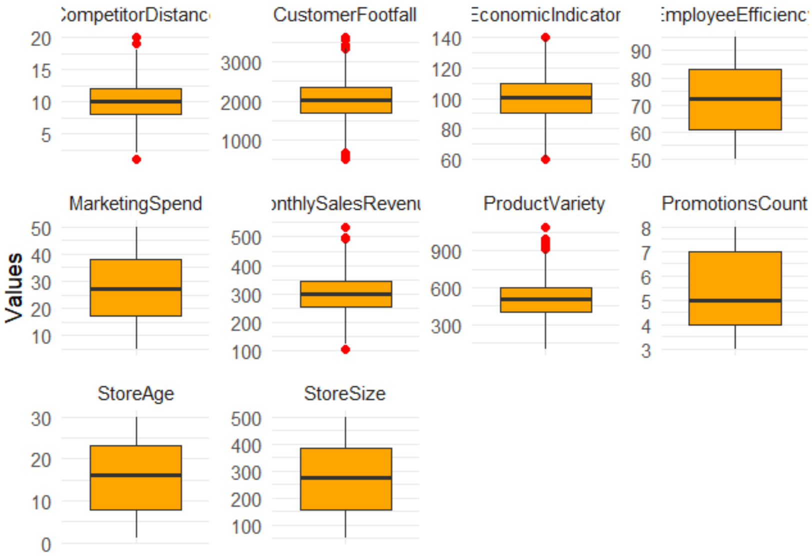

Figure 3.3: Boxplot Graphics of Variables

Although boxplot analysis identified a limited number of statistically extreme observations, the skewness and kurtosis values indicate that the distributions of the variables are approximately symmetric and do not exhibit heavy tails. Therefore, these observations were not treated as problematic outliers. They were retained in the dataset, as they reflect genuine performance differences across stores rather than data anomalies.

3.3 Data Transforming

We prepared the data in two different ways to make sure both models work perfectly. This is because each model has different rules for handling categories

Table 3.1: Skewness ve Kurtosis Values of Variables

Variable	**Skewness**	**Kurtosis**
ProductVariety	0.064	-0.065
MarketingSpend	-0.003	-1.193
CustomerFootfall	0.017	-0.038
StoreSize	0.012	-1.216
EmployeeEfficiency	0.031	-1.180
StoreAge	0.001	-1.201
CompetitorDistance	0.005	-0.023
PromotionsCount	-0.005	-0.927
EconomicIndicator	-0.070	-0.123
MonthlySalesRevenue	0.083	-0.249

Thresholds: $|Skewness| > 1$ *and Kurtosis* > 3

and the size of numbers. By using these two methods, we followed the mathematical needs of each model type.

The categorical variables *StoreLocation* and *StoreCategory* were transformed using two different methods:

Linear and Tree Model (Regression, Random Forest): To prevent multicollinearity and avoid the *Dummy Variable Trap*, an $n - 1$ coding scheme was applied. For each categorical variable, one level was removed as a reference.

Clustering Models (K-Means): A complete One-Hot Encoding (n levels) scheme was retained. This ensures that distance calculations remain equidistant and prevents information loss for tree splitting criteria.

Table 3.2: One-Hot (n) vs. Dummy Coding ($n - 1$)

Original Value	**Encoding (n)**	**Encoding ($n - 1$)**
(StoreLocation)	*(For K-Means)*	*(For Regression/RF)*
Los Angeles	1	1
San Francisco	0	0
Reference City	0	*(Dropped)*

Note: In the $n - 1$ *approach, the reference category is Los Angeles for cities, and Clothing for store types.).*

3.4 Scaling

The algorithms we will use in the supervised and unsupervised learning phases (K-Means, Linear Regression, and Random Forest) work with data differently. In the current state of our dataset, there are significant scale differences between the variables. For example, "CustomerFootfall" takes values like 1500, while "Store Age" remains at small values like 3-5.

This situation particularly misleads algorithms that calculate mathematical distances and coefficients. The impact of a value of 1500 is overshadowed by the impact of a value of 3 due to the difference in numerical magnitude. To address this imbalance, specific preprocessing steps appropriate to the nature of each model have been applied:

K-Means and Linear Regression: Because these algorithms are sensitive to the distance and magnitude of data, scaling is necessary to ensure that variables are represented with equal weight. Otherwise, larger numbers are dominating impacts. Random Forest: Because it is decision tree-based and processes data by dividing it according to specific threshold values rather than its size, it can work with raw data without needing any scaling. So, any scaling is applied to Random Forest dataset.

Min-Max Scaling for Clustering: Min-Max Scaling is preferred in the K-Means algorithm.

This is because K-Means, being based on distance, can disrupt clusters of variables with very different standard deviations. The Min-Max method brings all variables into the same playing field without disrupting the original distribution structure of the data, and does not allow any single variable to dominate the cluster.

For Linear Regression and data leakage prevention normalization was applied to our regression model. However, a critical data leakage prevention measure was implemented: The dataset was first divided into 80 Training and 20 Test. The scaling rule was created using only the data from the Training set. This rule was then applied (transformed) to both the Training and Test sets.

3.5 Correlations and Multicolinearity

As a last step before starting modeling, we need to check the relation between our independent variables. If there is a high correlation between two or more variables, we need to drop one of it to avoid from multicollinearity problem in our regression model.

When we set a temporary basic model to see multicollinearity with

Table 3.3: Variance Inflation Factor (VIF) Results

Variable	VIF	Variable	VIF
PromotionsCount	7.530	ProductVariety	1.021
MarketingSpend	7.521	EconomicIndicator	1.012
StoreLoc. San Francisco	1.510	StoreAge	1.008
StoreLoc. Sacramento	1.497	EmployeeEfficiency	1.008
StoreLoc. Palo Alto	1.488	CompetitorDistance	1.008
StoreCategory. Electronics	1.362	CustomerFootfall	1.007
StoreCategory. Grocery	1.360	StoreSize	1.006

Note: Threshold value is 5. (VIF > 5)

Variance Inflation Factor, we see that Marketing Spend and Promotion Count is highly correlated. It also says that most of the share of marketing spending is about promotions itself. So, we can drop promotion count from our dataset and not use in the models.

Table 3.4: Recalculated VIF Results (Top 4 Variables)

Variable	VIF	Variable	VIF
StoreLoc. San Francisco	1.509	StoreLoc. Palo Alto	1.488
StoreLoc. Sacramento	1.497	StoreCat. Electronics	1.362

Note: All independent variables are ¡ 5.

Chapter 4

Unsupervised Learning

4.1 K Means Clustering

The K-Means Clustering algorithm is used to discover natural groups within the data. The main objective of this analysis is to divide each store in the dataset into clusters based on the mathematical similarity of its attributes (number of visitors, store age, sales performance, etc.). In this way, we aim to identify different segments of stores which are exhibiting similar behaviors, creating a statistical basis for making strategically decisions specific to each group.

4.1.1 Elbow Method and Silhouette for Optimal K

Before running the K-Means algorithm, we need to decide how many groups the data should be ideally divided into. Instead of randomly choosing this number, two different graphical methods were used to find the ”K” value that best suits the structure of the data:

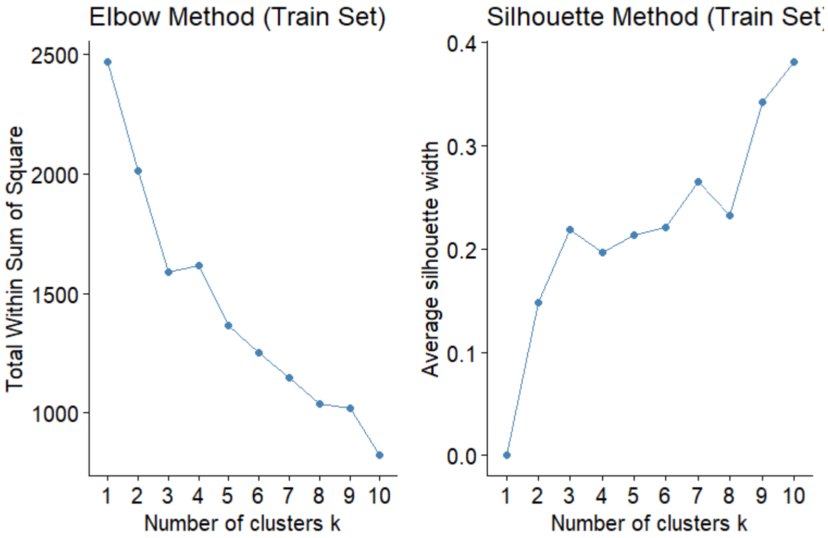

Figure 4.1: K-Elbow (left) and Silhouette Methods (right)

Elbow Method shows where increasing the number of groups reduces the contribution to the model. Examining our graph on the left, it can be seen that the point where the sharp drop of the curve ends and begins to flatten at point 3. After this point, increasing the number of groups does not make a significant difference.

Silhouette Method measures how clearly and smoothly the created groups are separated from each other. In the right graph, a distinct peak point is seen in the 3-cluster option. This confirms that the data is best divided into 3 parts without mixing. Although the 9th and 10th cluster options show a high result in the graph, it was not preferred because it is difficult and complex to manage and risk of overfitting. That's why, we are considering k as 3 because of lowest peak point, and continuous decrease after it.

4.2 Creating Clusters

Following data preprocessing steps, the optimal number of clusters was determined as k=3 using Elbow and Silhouette methods, and the initial model was established. When the resulting clustering was visualized, it was observed that the clusters were separated from each other by sharp boundaries.

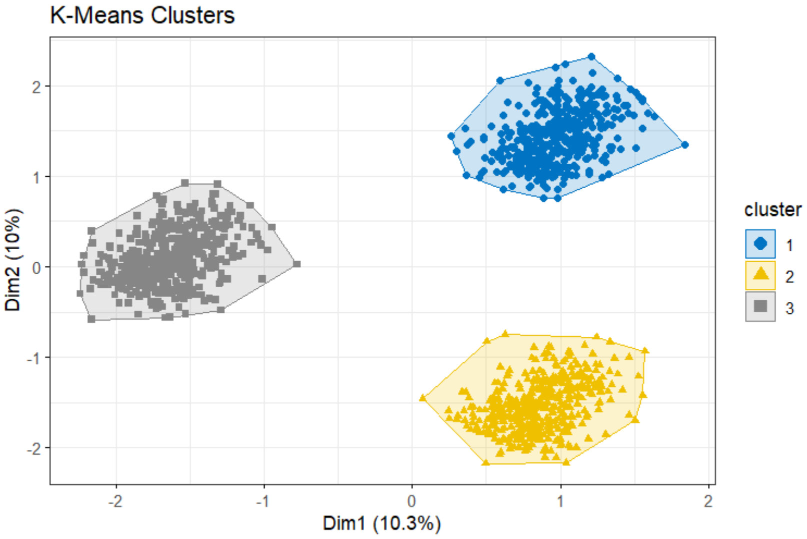

Figure 4.2: k=3 Clustering with Dummies

However, upon examination of the data, it was determined that this separation was predominantly driven by categorical variables rather than store performance. The algorithm grouped stores directly according to Clothing, Electronics, and Grocery categories, rather than their turnover. This was removed from the regression model because it would be a repetition of the category information already present in the model.

The model was revised to focus segmentation on performance metrics such as ”Volume,” ”Revenue,” and ”Physical Attributes,” regardless of store type. Categorical variables were excluded from the analysis, and the model was retrained using only numerical variables.

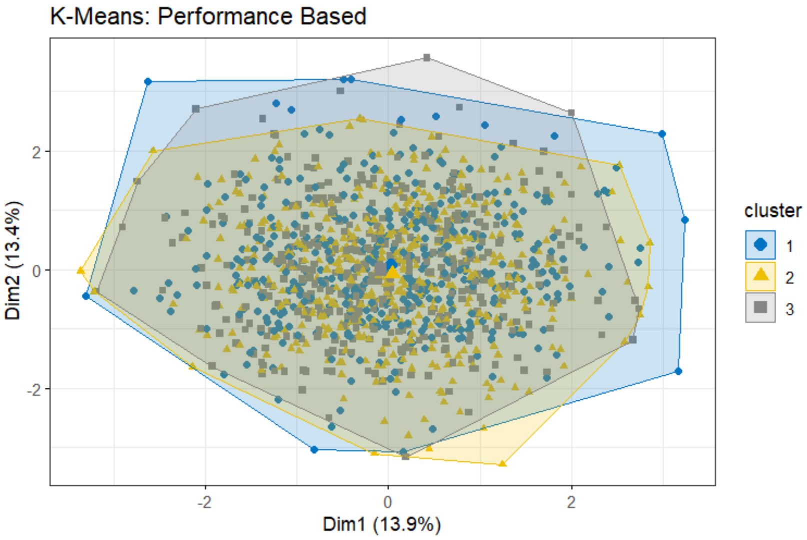

Figure 4.3: k=3 Clustering without Dummies

Visualizing the improved model showed that the clusters overlap and are closer together, instead of being completely separate like before. This suggests that the performance data (such as Sales and Store Size) changes gradually rather than having sharp limits. When looking at the statistics, we found no major differences between the three groups in key areas like ’Monthly Sales Revenue.’ Overall, the stores show very similar and consistent performance.

Chapter 5

Supervised Learning

5.1 Linear Regression

The linear regression model was developed to identify factors affecting store performance and to estimate monthly sales revenue. In the preprocessing part, we already split our data train-test and Z-Score Standardization was applied to them for interpreting in terms of effect size, regardless of the units of the variables. Also, we include clusters which we created with K-Means before to our model to see impact of them.

Table 5.1: Final Regression Results

Variable	**Estimate**	**Pr(>|t|)**
(Intercept)	296.748	¡ 2e-16 ***
ProductVariety	43.879	¡ 2e-16 ***
MarketingSpend	0.373	0.640
CustomerFootfall	-0.883	0.269
StoreSize	38.439	¡ 2e-16 ***
EmployeeEfficiency	3.585	7.63e-06 ***
StoreAge	1.086	0.174
CompetitorDistance	-0.130	0.871
EconomicIndicator	0.460	0.565
StoreLoc. Palo Alto	3.155	0.169
StoreLoc. Sacramento	0.975	0.662
StoreLoc. San Francisco	2.821	0.205
StoreCat. Electronics	-0.560	0.778
StoreCat. Grocery	0.020	0.992
cluster2	1.793	0.356
cluster3	1.468	0.464
Adjusted R-sq:	0.8049	

In the results of our linear regression model, we see that ProductVariety, StoreSize and EmployeeEfficiency variables are statistically significant and have impact on our model. A 1-unit increase in product variety cause an increase of approximately 43.88 units in sales. Store Size is the second strongest factor influencing sales. As store size increases, sales increase significantly, with 38.44 units. Employee Efficiency also has a positive and significant effect on sales. Location and Category does not play a distinctive role in sales performance.

These statistical results show that in-store operations are more important for success than external factors. The analysis proves that Product Variety is the strongest factor; customers prefer stores with many choices, and having a wide range of products on the shelves increases revenue more than advertising (Marketing Spend). Store Size is the second most important factor, showing that physical space is still necessary for retail success. However, the low correlation between size and variety ($r = 0.01$) suggests that true efficiency comes from using a large space to offer the widest possible product range. Furthermore, the positive impact of Employee Efficiency confirms that high quality service from staff leads directly to higher sales. Finally, because variables like location, category, and Customer Footfall were not significant, we can conclude that physical location and the size of the crowd do not limit revenue if the store manages its products and space effectively.

The high p-values of cluster2 and cluster3 indicate that the groups from the clustering analysis do not create a significant difference in sales compared to the based cluster 1.

Also, the model has normal residiual distributions and no heteroskedasticity problem.

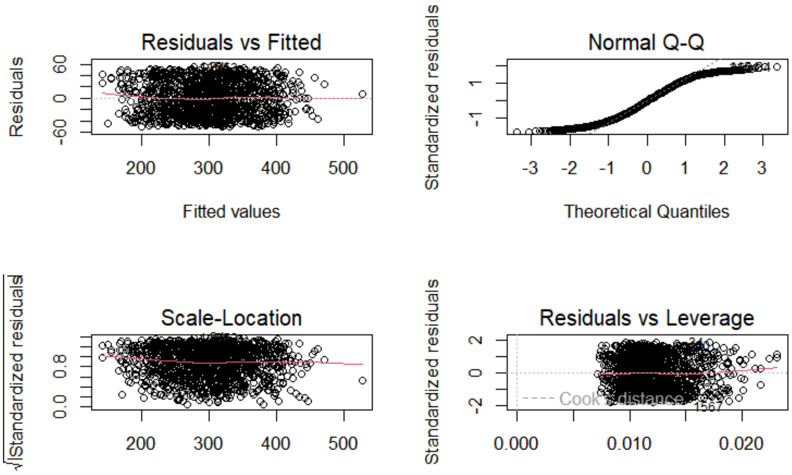

Figure 5.1: Q-Q Plot for Residuals Distribution

Table 5.2: Breusch-Pagan Test for Heteroskedasticity

Test	BP Statistic	df	p-value
Breusch-Pagan	17.211	15	0.3064

Q-Q Plot and Breusch-Pagan test result shows that the model has normal residiuals distributions and no heteroskedasticity problem.

Table 5.3: Model Performance Metrics

RMSE	R-squared	MAE
28.138	0.820	24.074

Our model's test results show that it is very successful at making predictions, with an R-squared of 0.819 and an RMSE of 28.14. Results shows that the model learned the patterns correctly and did not just memorize the data (overfitting). Therefore, we can safely use this model to make decisions for opening new stores or analyzing performance. Also, the low error rate (MAE: 24.07) shows that the model gives very accurate results for business planning.

5.2 Random Forest

The Random Forest model was developed to capture complex and nonlinear relationships affecting sales revenue. During model setup, cluster information obtained from K-Means clustering analysis was added as a new feature to the model. To ensure the most reliable results, 5-fold cross-validation was applied, and as a result of hyperparameter optimization, the structure where 10 variables are tested at each decision tree node (mtry = 15) was selected as the most optimal model. Because it is a tree-based algorithm, standardization was not applied to preserve the original units of the variables; the model proceeded with the raw data.

Table 5.4: Model Performance Metrics (Validation Set)

RMSE	**R-squared**	**MAE**
29.798	0.798	25.279

The developed Random Forest model performed strong on the test data, explaining the variance in sales with an accuracy of over 80 percent, with an R-squared value of 0.805. The fact that the margin of error in the predictions (RMSE: 29.422) parallels the linear regression results proves that the model has a stable structure and learns from the data without overfitting. Variable Importance analysis revealed that the main drivers of sales are Product Variety (100.00) and Store Size (99.9) by a significant margin. These results scientifically confirm that in-store capacity and product variety play the most critical role in sales success.

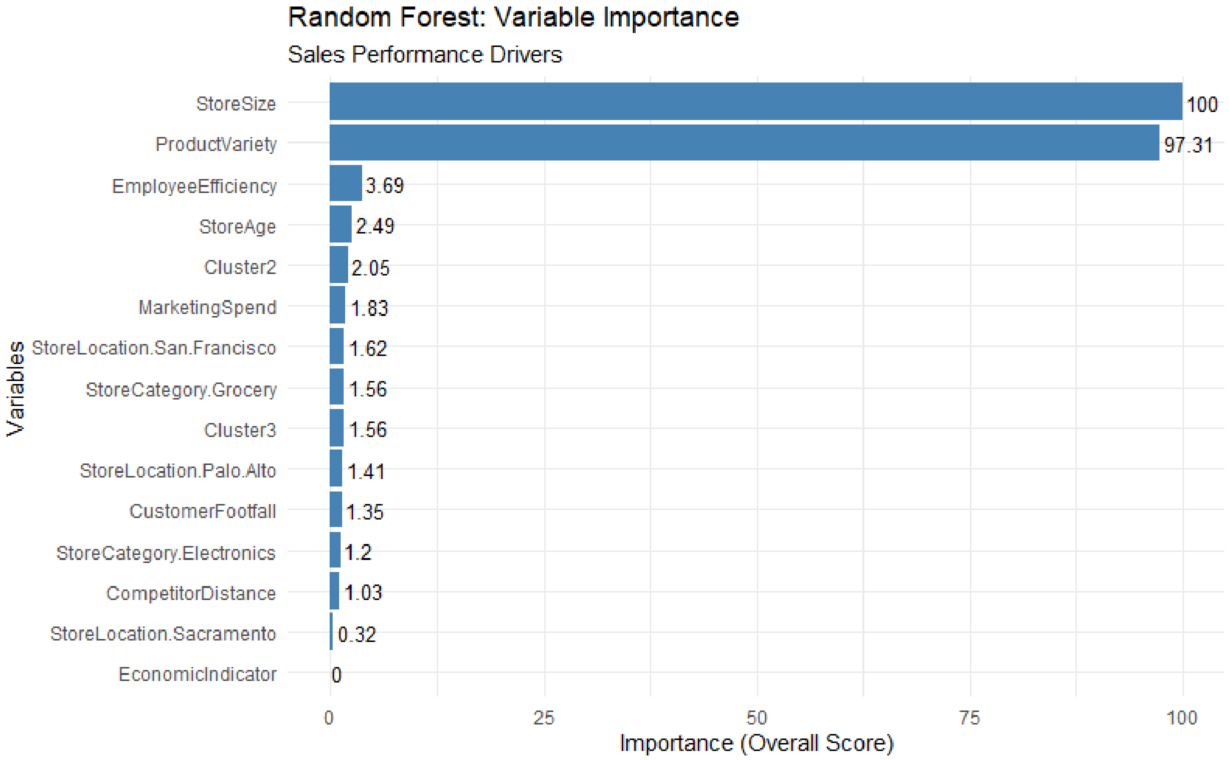

Figure 5.2: Feature Importance of Random Forest Model

Chapter 6

Conclusion

In this study, a retail store dataset with 1650 data points are analyzed with statistical methods to identify factors which are impactful on store revenues. In the study, both Unsupervised (K-Means) and Supervised (Linear Reg. , Random Forest) methodologies are combined and used, and created outcomes about how store attirbutes are shaping the success of a retail store.

The analysis results showed that both models have predictive capacity. The Linear Regression model perfectly explained the linear relationships between the variables with its R-square value of 0.819 and RMSE value of 28.14 obtained in the test set. As a result of experiements, it is found that Product Variety and Store Size factors are the most crucial for financial success. While Random Forest is giving highest importance values to these two factors (100, 97.3), marketing spendings or geographical locations were performing very low. Additionally, RF model has 0.798 R-square value, which means high explainability to our dataset, and identified variables are really strong predictors for our target variable.

From a management perspective, the findings suggest that strategic investments should prioritize inventory diversification and optimization of physical store space, rather than aggressive marketing campaigns or location-based targeting. Segmenting stores into three performance clusters via K-Means provides a practical framework for managers to tailor operational strategies to specific store profiles. However, because of the nature of our dataset, there was not so seperated clusters and there was not many different between 3 of clusters.